University of Zurich[UZH]

Speaker adaptations as a function of message, channel and listener variability

Thesis (cumulative thesis)
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by

Omnia Ahmed Mohamed Abdo Ibrahim

Supervisory Committee: Prof. Dr. Volker Dellwo (principal supervisor), Prof. Dr. Bistra Andreeva and Prof. Dr. Bernd Möbius

Accepted in the fall semester 2022 on the recommendation of the doctoral committee composed of Prof. Dr. Alexis Hervais-Adelman and Prof. Dr. Petra Wagner

Zurich, 2023

Acknowledgment

I would like to thank

My supervisors: Prof. Dr. Volker Dellwo, Prof. Dr. Bistra Andreeva and Prof. Dr. Bernd Möbius

for guiding my way and supporting me throughout this process.

Dr Ivan Yuen and Prof. Dr. Gabriel Skantze

without whom some of this work would not have been possible.

My master supervisor: Prof. Dr. Mervat Fashal

for her continuous valuable support at various levels beyond and after my master degree.

Raphael Werner, Beeke Muhlack, Mikey Elmers and Dr. Iona Gessinger

for valuable support at various levels

My brothers

for everything.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). Project 232722074: SFB 1102 Information Density and Linguistic Encoding. I was lucky to be part of this SFB as a Phd student in Project C1 'Information Density and the Predictability of Phonetic Structure'.

The first year of my PhD was funded by the University Research Priority Program (URPP), “Language and Space" lab, University of Zürich.

Abstract

Speech is a highly dynamic process. Some variability is inherited directly from the language itself, while other variability stems from adapting to the surrounding environment or interlocutor. This Ph.D. thesis consists of seven studies investigating speech adaptation concerning the message, channel, and listener variability. It starts with investigating speakers' adaptation to the linguistic message. Previous work has shown that duration is shortened in more predictable contexts, and conversely lengthened in less predictable contexts. This pervasive predictability effect is well studied in multiple languages and linguistic levels. However, syllable level predictability has been generally overlooked so far. This thesis aims to fill that gap. It focuses on the effect of information-theoretic factors at both the syllable and segmental levels. Furthermore, it found that the predictability effect is not uniform across all durational cues but is somewhat sensitive to the phonological relevance of a language-specific phonetic cue.

Speakers adapt not only to their message but also to the channel of transfer. For example, it is known that speakers modulate the characteristics of their speech and produce clear speech in response to background noise—syllables in noise have a longer duration, with higher average intensity, larger intensity range, and higher F0. Hence, speakers choose redundant multi-dimensional acoustic modifications to make their voices more salient and detectable in a noisy environment. This Ph.D. thesis provides new insights into speakers' adaptation to noise and predictability on the acoustic realizations of syllables in German; showing that the speakers' response to background noise is independent of syllable predictability.

Regarding speaker-to-listener adaptations, this thesis finds that speech variability is not necessarily a function of the interaction's duration. Instead, speakers constantly position themselves concerning the ongoing social interaction. Indeed, speakers' cooperation during the discussion would lead to a higher convergence behavior. Moreover, interpersonal power dynamics between interlocutors were found to serve as a predictor for accommodation behavior. This adaptation holds for both human-human interaction and human-robot interaction. In an ecological validity study, speakers changed their voice depending on whether they were addressing a human or a robot. Those findings align with previous studies on robot-directed speech and confirm that this difference also holds when the conversations are more natural and spontaneous.

The results of this thesis provide compelling evidence that speech adaptation is socially motivated and, to some extent, consciously controlled by the speaker. These findings have implications for including environment-based and listener-based formulations in speech production models along with message-based formulations. Furthermore, this thesis aims to advance our understanding of verbal and non-verbal behavior mechanisms for social communication. Finally, it contributes to the broader literature on information-theoretical factors and accommodation effects on speakers' acoustic realization.

Contents

1 Synopsis **1**

1.1 Motivation 1

1.2 Theoretical frameworks 6

1.2.1 Information theory 6

1.2.2 Hypo & Hyper articulation model 8

1.2.3 Accommodation theory 9

1.3 Thesis outline 11

I Content-driven adaptation **16**

2 The effect of predictability on German stop voicing is phonologically selective **17**

3 Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles **18**

II Acoustic noise-driven adaptation **19**

4 The effect of Lombard speech modifications in different information density contexts **20**

5 The combined effects of contextual predictability and noise on the acoustic realization of German syllables **21**

III Listener-driven adaptation **22**

6 Within-speaker accommodation behavior in apology-centered interactions: The role of socio-pragmatic factors **23**

7 Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing **25**

8 Revisiting robot directed speech effects in spontaneous Human-Human-Robot interactions **26**

9 Future directions **27**

9.1 The interplay between predictability and speech tempo effects on Modern Standard Arabic 27

9.2 The relationship between predictability and accommodation behavior 28

9.3 What makes a good counsellor? Using phonetic accommodation to distinguish between high-quality and low-quality counselling conversations 29

Bibliography **31**

Chapter 1

Synopsis

"Adaptability is not imitation. It means power of resistance and assimilation."

Mahatma Gandhi

Communication is one of the most basic human needs. In everyday communication, the speaker aims to communicate their messages intelligibly to their listeners. Therefore, when they are aware of any speech perception difficulty on the part of the listener due to background noise, a hearing impairment, or a different native language, speakers will naturally and spontaneously modify their speech to accommodate the listener (Bradlow, 2002).

The current thesis explores speaker adaptation with the aim of better understanding human voice dynamics beyond the existing literature. The thesis consists of seven studies investigating speech variability as a function of I) message encoding, II) channel encoding, and III) listener encoding. The findings of this thesis could advance our understanding of the mechanisms underlying verbal and non-verbal behavior for social communication and contribute to the broader literature on information theoretical factors effects and accommodation by investigating different languages (Arabic, German, Catalan, Swedish). This chapter will provide an overview of the thesis by first discussing the motivation, followed by the theoretical framework and outlining the various studies.

1.1 Motivation

In September 2019, during Interspeech in Graz, Furhat Robotics (a robot company) offered to play a sorting-cards game with the robot in its exhibition. During the noisy conference break (Figure 1.1), I started the game with a friend. It was genuinely interesting how I simultaneously adapted to the noisy environment and my interlocutors. I aimed to help my listeners (friend and robot) without exhausting my voice. I spoke

Figure 1.1: The speech processing community during one of Interspeech2019's coffee break

clearly with the robot and hyperarticulated every word with longer pauses between words. With my friend- it was less extreme- I was getting closer to her to reduce the distance so she could hear me. While with the salesman, who was at a distance, I used more gestures accompanying my speech. So in that short conversation, I was juggling three communication strategies; clear speech, reduce the distance and combined gestures. Consequently, my adaptation to the background noise was different for each interlocutor.

Generally, speakers exhibit a great deal of acoustic-phonetic variability due to physiological differences (e.g., age, gender, and vocal track size) and psychological states (e.g., emotions, stress). Furthermore, in everyday life, speech dynamically varies according to different factors such as situation (e.g., formal vs. informal), function (e.g., request or apology), topic, and interlocutors, amongst others (Leongómez et al., 2021). All of these factors lead to within-speaker variability and are known to serve a communicative purpose and have perceptual consequences (see Cooke et al. (2014) for a review). This Ph.D. thesis investigates, through various experiments, how speakers control and improve their speech production so that communication runs smoothly.

Within-speaker variability could be inherited directly from the language itself. For example, speakers tend to reserve effort while aiming for efficient communication. They produce more reduced forms or shorter durations for predictable/probable messages. This predictability effect reflects a tendency towards communicative efficiency when the sender and receiver can achieve successful communication with minimal effort on average. A growing body of research suggests the pervasiveness of predictability in language, as manifested in the phonetic reduction of words with high probability (Aylett & Turk, 2006; Pate & Goldwater, 2015; Aylett & Turk, 2004), or choices of linguistic/syntactic units (Jaeger, 2010). However, it is still not clear how

predictability effects are transmitted across different linguistic levels.

Speakers adapt not only to their message but also to the channel of transfer. For example, it is known that exposure to background noise will trigger speaker adaptation, also known as 'Lombard speech' (Lombard, 1911; Brumm & Zollinger, 2011). Its goal is to enhance communication in challenging and degraded interactive situations. Such speech adaptation often leads to increased loudness, higher pitch, expanded vowel space, hyper-articulation and lengthening (Castellanos et al., 1996; Boril & Pollák, 2005; Lu & Cooke, 2009b), with perceptual consequences of improved intelligibility (Garnier & Henrich, 2014; Hansen et al., 2020). Lu & Cooke (2008) found that Lombard speech is more intelligible than 'plain' speech if presented in a noisy background. However, Lombard speech modification does not manifest uniformly across different segment types or linguistic units. Instead, Lombard speech occurs in units that are critical to intelligibility. For instance, speakers emphasize vowels more than consonants in a noisy environment; presumably, the former is more critical to speech audibility (Garnier & Henrich, 2014). Therefore, speakers would likely enhance information carrier units when speaking in noise. Therefore, this thesis investigates further the combined effects of noise and predictability.

Speech adaptations are not only motivated by the improvement of speech intelligibility but also by social goals. When communicating, speakers constantly position themselves concerning ongoing social interaction. Depending on whether a person interacts with a senior or a peer, the socio-pragmatic qualities of their speech is adapted in one way or another (Winter et al., 2021). This adaptation may lead to an increase or a decrease in the interlocutors' speech similarity, known as accommodation/alignment. There is mixed evidence on whether accommodation between interlocutors increases in a linear fashion throughout the course of an interaction or whether it has dynamic characteristics. The direction and extent of phonetic accommodation have been found to depend on factors such as the attitude or the hierarchical relationship towards an interlocutor. For example, it has been shown that an increase in the likability of a conversational partner led to a stronger convergence effect for vowel quality (Schweitzer & Lewandowski, 2014) and fundamental frequency (Michalsky & Schoormann, 2017). Furthermore, Gregory Jr. & Webster (1996) suggests that speakers on the lower end of the social hierarchy, or in a less dominant role, converge towards the style of the hierarchically higher or more dominant interlocutor. Those findings suggest that accommodating behavior is socially motivated and, to some extent, controlled by the speaker.

This speaker-to-listener adaptation extends to human-robot interactions. With the need to communicate effectively with computers or robots, speakers have been shown to address computers/robots differently than humans. Speakers change their acoustic characteristics (ex: raise their F0, slow their speaking rate) when they are talking to a robot in comparison to a person (Kriz et al., 2009). However, nowadays, robots have become more and more part of our life. Many people interact with robots in different ways; for example, in health care, robots could help people who need special

care, like people with dementia, or help lift people in a hospital. Robots could also serve social purposes like helping in education or providing information in a train station. With the advancement of AI-powered speech technology, robots show improved performance. For example, Google announced in 2017 that its speech recognition is almost as accurate as humans. However, they are concerned that their speech synthesis technology that creates human-like voices could be abused with Deep-fake. Despite, their performance are catered for specific situations and languages, but in general, we cannot neglect the tremendous technological advances. It thus seems plausible that such advances could reduce the well-known effect of robot-directed speech in a real-life setting.

Human-human interaction can be a useful starting point for designing robot-human interactions, but it is not necessarily a perfect or complete model. For example, studies on how people interact in different contexts, such as meetings, interviews, or social gatherings, can help develop robots that adapt to these contexts and effectively engage with humans. Therefore while robots can undoubtedly learn from how humans communicate and interact with each other, there are also significant differences in how humans and robots operate. For example, humans can recognize and respond to complex emotional cues in ways that state-of-the-art robots cannot replicate. Humans can also understand a given interaction's social and cultural context, which is not always easy for robots to do. Furthermore, humans are often more able to change and adapt to unexpected situations during conversations, while robots may struggle when presented with novel scenarios or information. Moreover, robots may have different limitations and abilities than humans, requiring unique approaches to designing their interactions with humans. Thus, while human-human interaction can provide essential insights for robot-human interactions, it is crucial to consider robots' unique characteristics and limitations when designing their interactions with humans. I will explore in this thesis the extent to which humans adapt their speaking style to the listener in unstructured human-human-robot interactions (chapters 7 and 8).

Speaker adaptation has multiple sources (Figure 1.2). The subdivision into message-, environment- and listener- related adaptation is by no means absolute (Cooke et al., 2014). They tend to co-occur together in the same conversation. This is Similar to my experience during Interspeech 2019, where I adapted to different listeners (human and robot) in noise. Generally, in our everyday speech, communication rarely happens in a noise-free environment; speakers often need to adapt their ways of talking to counter noise (e.g., in noisy restaurants) to ensure listeners correctly understand their messages. To date, the process underlying how speakers adapt when different communication demands co-occur and/or conflict is poorly understood. For example, speakers enrich their speech in the presence of noise; however, they also modify their speech to be efficient by shortening word duration in more predictable contexts. To meet these two communicative functions, speakers will attempt to resolve these conflicting demands. How do speakers estimate and weigh the needs of their listener(s) and their own? Will the adaptations show an additive effect (e.g., unpredicted words

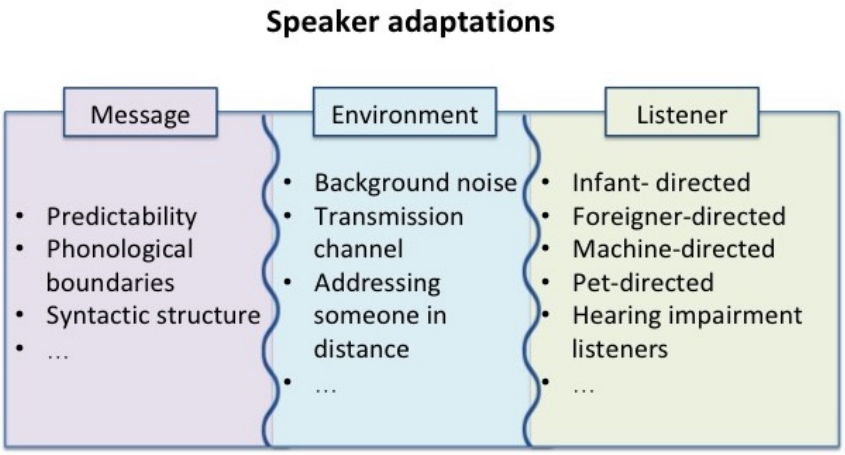

Figure 1.2: Different sources of speaker adaptations

will have a longer duration, and background noise will add further increase), or will they interact? The present thesis aims to fill this gap by investigating the speaker adaptation to linguistic message and noise (chapters 4 and 5). Then, how do speakers express an apology speech act while adapting to listeners with different power dynamics (chapter 6).

Moreover, speech and language sciences aim to understand human behavior in the real world. However, a large proportion of existing work on speaker adaptations utilizes laboratory settings, where speakers are aware that their speech is recorded and likely to adopt a formal speaking style of what is known as the observer's paradox (Labov, 2006). This makes conventional lab setting results hard to reproduce in the real world and their results can not be generalized. It is important to explore how lab results are applied to real-life situations outside a controlled setting. To bridge the gap between lab and real-life, this thesis attempts to revisit some of speaker adaptation issues in unscripted real-life scenarios, viz. interactions in a science museum (chapters 7 and 8).

This thesis investigates different aspects of within-speaker adaptations to the linguistic message, the environment, and the listener. It specifically addresses the following questions:

- To what extent do speakers change their voice depending on their message, environment or who they are talking to?
- What will happen if two sources of speaker adaptations co-occur together?
- How does a naturalistic setting change speaker adaptations?
- Which factors drive those adaptations in Human-Human and Human-Robot interaction?

One of the implication of my work is to understand human-human interaction better so that we can transfer our knowledge to human-robot interaction. It is challenging because we try to model human experience and behavior in machines. It is necessary to move human-robot interactions into more naturalistic and real-life scenarios. For example, despite the vast advances in technology, automatic speech/speaker recognition systems still face many challenges in spontaneous speech. A better understanding of speech variability will help enhance system performance. Furthermore, the study of within-speaker variability is of particular interest to forensic phonetics since samples in forensic speaker comparisons tend to be mismatched in speaking situation and style.

1.2 Theoretical frameworks

This thesis contributes to a diverse range of phonetic studies investigating variability as a function of the message, channel, and listener variability. Before introducing the conducted studies, firstly, I will introduce three theories, which help give insights into speaker adaptations.

1.2.1 Information theory

"You should call it entropy, for two reasons. In the first place, your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate, you will always have the advantage."

John von Neumann
Suggestion of von Neumann to Shannon
regarding the name of his new uncertainty function

Information theory studies the quantification, storage, and communication of information. Claude E. Shannon initially proposed it in 1948 to find fundamental limits on signal processing and communication operations in a landmark paper entitled "A Mathematical Theory of Communication". The foundations of the discipline were laid down by Shannon (1948) and then extended by Warren Weaver (Shannon & Weaver, 1949), who established the theory that found the applications far beyond the initially intended fields of cryptography and telecommunication.

The goal of the information-theoretical model of information exchange firstly proposed to optimize communication between machines has been defined as maximizing the amount of information transferred via a channel while considering its intrinsic limitations. According to Shannon, "information" is thought of as a set of possible messages, where the goal is to send these messages over a noisy channel, and then to

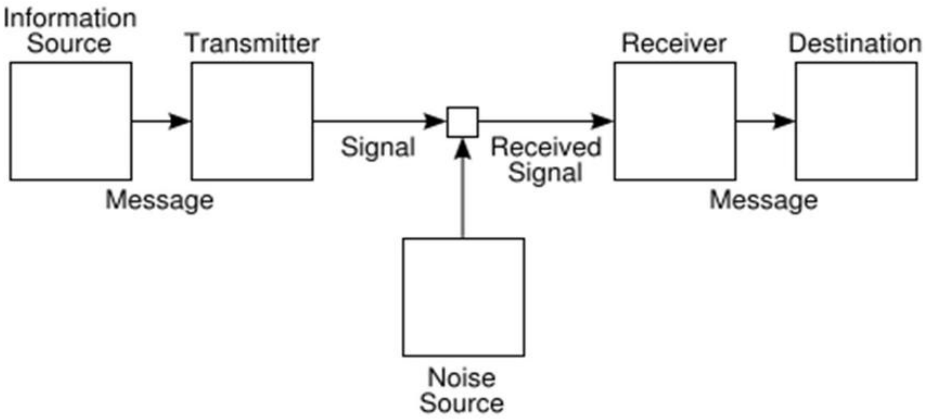

Figure 1.3: A schematic diagram of Shannon (1948)'s model of communication

have the receiver reconstruct the message with a low probability of error, despite the channel noise (Figure 1.3).

The concepts of information theory have been adapted in many scientific fields, among those in linguistics. Speakers share knowledge of language-specific probability distribution of linguistic units. Therefore, they tend to produce more reduced forms or shorter durations for predictable/probable messages, known as the predictability effect. Not every letter of a word carries the same amount of information. Fore example, in English, after a "th", the following letter most probably will be a "e" to form one of the most frequent English word "the" (van Heuven et al., 2014). So "e" here in this context carries very little information because you could predict it beforehand and could consider it redundant. That is why we can make sense of things like this: "if u cn rd ths u cn gt a gd jb"[1].

Predictability is known to facilitate speech perception in challenging listening conditions (ex: background noise). High-predicted words are easier to perceive in a noisy environment. However, it can lead to mishearing - high confidence in incorrect responses based on context (Van Os et al., 2021). Furthermore, Michael & Ibrahim (2022) found that listeners could recover from missing words resulting from packet loss if the words were highly-frequent. The packet loss-affected speech cannot be modeled with parameters of the transmission network alone but needs information about the importance of the lost signal.

In recent years, information-theoretic principles have helped gain insight into the predictability effect on linguistic variability. Studies have found that linguistic units are prone to reduction or deletion when they are easily predictable from their context. Its effect is quite common and general, as evident in a survey of 600 languages (Pimentel et al., 2021). These findings hold at discourse (Torabi Asr & Demberg, 2015), sentence (Jaeger, 2010), word (Buz & Jaeger, 2016), syllable (Aylett & Turk, 2006; Ibrahim et al., 2021), or phoneme (Aylett & Turk, 2006; Bybee, 2002) levels. Gener-

[1] "If you can read this you can get a good job

ally, duration is shortened in more predictable context and conversely lengthened in less predictable context. It has been suggested that long duration results from explicit encoding to improve the intelligibility of hard-to-understand units (Jaeger, 2010; Gahl et al., 2012). So far, little attention has been paid to syllable level predictability. This thesis aims to fill this gap. It focuses on the effect of information-theoretic factors on both the syllable and segmental levels.

Several measures have been proposed to objectively assess the amount of information transferred via a known channel at a given time (Brandt et al., 2019; Hale, 2016). Metrics, such as surprisal and entropy, have often been used to answer the question of linguistic or acoustic redundancies driven by efficient communication principles and optimal use of channel capacity. Surprisal, or self-information, is an information-theoretic measure that quantifies the (un)expectedness of a particular outcome, measured in bits, inversely proportional to its probability (Shannon, 1948). It is relevant for human processing difficulty of linguistic units at different levels (Hale, 2001; Demberg et al., 2012) More specifically, surprisal is calculated as the negative logarithm (base 2) of the probability of any linguistic unit to occur using (see Equation 1.1). From this equation, units that are less predictable score high surprisal values, and vice versa.

$$S(unit_i) = -log_2 P(unit_i|Context) \tag{1.1}$$

In the current thesis, a syllable-based language model is used to estimate the probability of a unit given its previous local context, based on large text corpora. Therefore, surprisal is defined as the negative probability of a syllable to occur given its two preceding syllables.

1.2.2 Hypo & Hyper articulation model

Speech production is highly adaptive. Speakers can and typically do tune their language to meet communicative and situational demands. Goal-oriented speaking styles, such as infant- or robot-directed speech, can be seen as an adjustment of the speaker's output to meet the demands of their target audience or the communicative situation (see Cooke et al. (2014) for a review). This flexibility is backed up by the hyper-and hypo-articulation (H&H) theory, which proposes that speakers can vary their speech output along a continuum of hyper-speech and hypo-speech to strike a balance between speaking as clearly as possible for the sake of the listener (hyper-articulated speech), while spending as little effort as possible (hypo-articulated speech) (Lindblom, 1990). The cause of this variability is the speaker's adaptiveness, where s/he balances the listener's need for a clear signal with energy preservation. Those within-speaker variations are a reflection of the trade-off between clarity of speech (listener-oriented output) and economy of effort (talker-oriented) (Turnbull, 2019). To prevent speakers from over-economising to the point of unintelligibility, hypoarticulation is usually governed by a constraint of linguistic distinctiveness; speak-

ers hypoarticulate only while listeners are able to distinguish the target from competing lexical items.

Hypoarticulated speech is characterized by a lower fundamental frequency (F0), spectral tilting, shorter duration (Rouas et al., 2010), and smaller vowel space (Englund, 2017). These acoustic properties often make an acoustic signal phonemically confusable even for a human. At the other end, hyperarticulation is defined as increasing the articulatory efforts to maximize the clarity of speech. The advantage of these properties is to make acoustic speech more intelligible in human communication. However, the speech variability caused by hyperarticulation is not always good for automatic speech recognition systems since those systems are primarily modeled on casual speech data.

There is a connection between H&H theory and the predictability effect; if the speaker has reasons to assume that the listener can predict what will come next, s/he tends to hypoarticulate and reduce the signal, whereas more unexpected content is more likely to be hyperarticulated (Munson & Solomon, 2004). Chapters 3, 4 and 5 investigate the effect of predictability on German syllables. Those chapters also investigate the Lombard effect. Talkers involuntarily produce more intelligible speech when they are in a noisy environment (Lombard, 1911). Such speech provides more acoustic redundancy, such as greater amplitude, slower speech, and more energy at higher frequencies, precisely when the received acoustic signal will be more ambiguous (see Junqua (1996) for a review). Previous work on the Lombard effect shows that talkers guard against degraded channels (consciously or not); when talkers hear noise, they produce louder, more intelligible speech.

Finally, based on the H&H theory, it is reasonable to hypothesize that the speaker will adapt to the limited understanding capabilities of the robot resulting in hyperarticulation in robot-directed speech, which is investigated in chapter 8 in human-human-robot setting.

1.2.3 Accommodation theory

Accommodation - the process by which interlocutors change their speech patterns either by becoming more similar (convergence) or dissimilar (divergence) - is a pervasive phenomenon in speech communication (Giles et al., 1991). It elaborates on the human tendency to adjust their behavior while interacting. The accommodation theory assumes interpersonal conversation to be a dynamic adaptive exchange of verbal and nonverbal behavior. The tendency for interlocutors to become more similar in their speaking behavior has been referred to in various disciplines as convergence (Pardo, 2006), entrainment (Levitan & Hirschberg, 2011), alignment (Pickering & Garrod, 2004, 2013, 2021), accommodation (Giles et al., 1991; Shepard et al., 2001), or adaptation (Bell et al., 2003). While all of these terms refer to aspects of how speakers adjust their speech patterns in conversation, they each highlight different aspects of the process and may reflect different motivations or goals. In this thesis, I

will use the term “accommodation” to describe the phenomena of speakers adjusting their speech during the interaction, while the terms “convergence" and “divergence” defined as the phenomenon of interlocutors’ speech becoming more or less similar during an interaction.

Many factors affect phonetic convergence. For instance, less convergence has been documented in speakers with high-self rated autonomy (Heath, 2017), and towards interlocutors perceived as more competent compared to those rated as more friendly (Schweitzer & Lewandowski, 2014). Furthermore, accommodation has been studied in association with dialogue quality, speech production and perception, participants’ trust, and interaction smoothness. In the current thesis, I investigate various social and situational factors which could affect phonetic accommodation.

Measuring accommodation could help assess the engagement dynamic of the conversation participants as an index to the success of the interaction. In addition, studying accommodation could help in understanding the nature of human (and also in animal kingdom (Ruch et al., 2018)) behavior in everyday interactions as a function of different social relations and norms (Zellers & Schweitzer, 2017) (interpersonal relationships), ex: power dominance – closeness – familiar/unfamiliar, etc. Namy et al. (2002) have found that female shadowers converged towards the model speaker to a greater extent than males. While more recent studies (Pardo, 2006; Babel, 2012; Pardo et al., 2017) showed inconsistency in that respect. In another study, Pardo et al. (2018) found no difference in the amount of phonetic convergence depending on speaker gender, whether in their conversational interaction task or speech shadowing task.

By better understanding accommodation, we could model convergence in human-computer (robot) interactions to increase engagement with the computer. Furthermore, we can build agents that interact more naturally by respecting principles of accommodation. Driven by its importance, various researchers develop accommodation measures based on local (turn by turn) and global (conversation level) events.

Accommodation is measured along two primary time scales referred to as local and global (Levitan & Hirschberg, 2011; Lee et al., 2014). In this thesis, I will focus on the global scale of accommodation at the conversation level. The traditional method of measuring accommodation is the difference of distance (DID), the change in the absolute distance between the interlocutors, the speaker and his interlocutor or model talker. There are two different recording tasks used for accommodation; shadow task and conversational interactions:

- Some accommodation studies measure similarity in particular acoustic characteristics using a shadow task, where speakers are exposed to recordings of a model talker, which they repeat after, and the comparison is made of their speech before and after exposure (Pardo, 2013; Babel, 2012).

- For more natural settings, researchers extract accommodation from the conversational interactions (at the beginning and end of conversation or different

turns). It presents additional complications in defining reference points for the speakers because both interlocutors are potentially changing (Pardo et al., 2018). I used conversational interactions in this thesis to measure accommodation (see chapters 5 & 6).

Their recent work Cohen Priva & Sanker (2019) demonstrates that DID is not a suitable measure of accommodation. One of its limitations is that it is based on comparing the distance between speakers' means without referring to the raw values for each speaker; DID measures require one single value to represent the acoustic feature, which is usually the mean value of the speaker's voice over a span of time. Other researchers used t-test, which determines if there is a significant difference between the means of two interlocutors' distributions (Levitan & Hirschberg, 2011). The described measures assume that the extracted acoustic values are normally distributed.

Unfortunately, the acoustic features are often not well described by normal distribution (Traunmüller & Eriksson, 1994; Simpson, 2009). The mean value has been proved that it is not a suitable measure for acoustic features, and then the mean is somehow misleading. For example, if F0 is scaled linearly (in Hz), there is some positive skewness (Mikeev, 1971). In addition, it has been observed that some speakers show a bimodal f0-distribution, in particular when speaking with increased vocal effort (Rappaport, 1958). In order to calculate DID for accommodation, we need to use a nonparametric test (distribution-free test), which does not assume that data are sampled from normal distribution.

Two possible nonparametric tests are the well-known Kolmogorov-Smirnov test (KS), which compares two distributions without comparing any particular parameter (i.e., mean or median), and the Anderson-Darling (AD) test is a modification of the KS test. It is used for the same purpose (Scholz & Zhu, 2013). However, they differ in that AD gives more weight to the tails of the distributions, while KS is more sensitive to deviations in the center of the distribution. The AD test was used to quantify the F0 modulation of different speaking styles (Arantes & Eriksson, 2019).

In the current thesis, I propose using the AD test to measure global level accommodation in conversational interaction. Unlike other existing methods, it does not compare the mean values from the interlocutors. However, It considers every single data point as it contributes to the overall shape of the distribution of the speaker's voice (see chapter 6).

1.3 Thesis outline

The following seven studies were conducted to meet those goals and to describe speakers' adaptations in four languages (Arabic, German, Catalan, and Swedish). Each study constitutes a separate chapter of this thesis.

Study 1: The effect of predictability on German stop voicing is phonologically selective

This thesis starts with investigating speakers' adaptation to the linguistic message. Cross-linguistic evidence suggests that syllables in predictable contexts have a shorter duration than in unpredictable contexts. However, it is unclear if predictability uniformly affects phonetic cues of a phonological feature in a segment. This study explored the effect of syllable-based predictability on the durational correlates of the phonological stop voicing contrast in German, viz. voice onset time (VOT) and closure duration (CD), using data in Ibrahim et al. (2021). I hypothesized that stops (voiced and voiceless) would have longer VOT and CD in less predictable syllables. The results showed an interaction effect of predictability and the voicing status of the target consonants on VOT, but a uniform effect on closure duration. This interaction effect on a primary cue like VOT indicates a selective effect of predictability on VOT, but not on CD. This suggests that the effect of predictability is sensitive to the phonological relevance of a language-specific phonetic cue and that syllable-based surprisal both directly and indirectly affects the phonological feature [voice], as it affects phonetic cues differentially, depending on how relevant and distinct each cue is to the phonological contrast in the language.

Study 2: Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles

In the second study, I collected an Arabic corpus in two different speaking styles, intending to investigate Arabic rhythm. It was motivated by the lack of an Arabic database for studying speech rhythm and tempo, while those databases exist for numerous languages. Previous Arabic rhythm studies mainly relied on either read or spontaneous data, while our corpus involves both speaking styles (read and spontaneous), which allows a better understanding of the Arabic rhythm.

10 Egyptian speakers (gender-balanced) produced speech in two different speaking styles (read and spontaneous). The design of the reading task replicates the methodology used in creating the BonnTempo corpus (BTC). During the spontaneous task, speakers talked freely for more than one minute about their daily life and/or studies. Then they described the directions to come to the university from a famous nearby location using a map as a visual stimulus (See Table 1.1). The database serves as a phonetic resource, which allows researchers to examine various aspects of Arabic supra-segmental features. It can be used for forensic phonetic research for comparing different speakers, analyzing variability in different speaking styles, and automatic speech and speaker recognition. In future work, this corpus will be used to investigate the interplay between predictability and speech tempo effects on Modern Standard Arabic (chapter 9)

Table 1.1: Content of Arabic Speech Rhythm Corpus

Type	Task	Duration	Total (X 10 speakers)
Read	Short story	40 minutes	400 minutes
Spontaneous	Interview questions	15 minutes	150 minutes
Spontaneous	Map (directions)	15 minutes	150 minutes

The Arabic corpus, along with other languages in the BonnTempo corpus, is available under request from prof. Dr. Volker Dellwo (volker.dellwo@uzh.ch)

Study 3 and 4: The combined effects of contextual predictability and noise on the acoustic realization of German syllables

In the 3rd and 4th studies, I moved to investigate channel related adaptation. Speakers adapt their speech to increase clarity in the presence of background noise (Lombard speech). However, they also modify their speech to be efficient by shortening word duration in more predictable contexts. To meet these two communicative functions, speakers have to attempt to resolve these conflicting communicative demands.

The 3rd study focuses on how this can be resolved in the acoustic domain. A subset of 1520 target CV syllables was analysed in two white-noise (reference = quiet vs. -10 dB SNR) and two Predictability (defined as surprisal) (H vs. L) contexts. The results revealed effects of both noise and surprisal on syllable duration and intensity range, but only an effect of noise on F0. This might suggest redundant (multi-dimensional) acoustic coding in Lombard speech modification, but not so in surprisal modification.

The 4th study investigates in more depth the acoustic realizations of syllables in predictable vs. unpredictable contexts across different background noise levels. Thirty-eight German native speakers produced 60 CV syllables in two predictability contexts in three noise conditions (reference = quiet, 0 dB, and -10 dB SNR). The presence of noise yielded significantly longer duration, higher average intensity, larger intensity range, and higher F0. Noise levels affected intensity (average and range) and F0. Low predictability syllables exhibited longer duration and larger intensity range. However, no interaction was found between noise and predictability. This suggests that noise-related modifications might be independent of predictability-related changes, with implications for including channel-based and message-based formulations in speech production.

The language models used in this thesis for syllables and phones, along with their description, are uploaded to CLARIN Virtual Language Observatory (VLO) and are available here.

Study 5: Within-speaker accommodation behavior in apology-centered interactions: The role of socio-pragmatic factors

Speakers modulate their speech and voice to match their interlocutors' characteristics. Studies have shown that convergent vocal accommodation (i.e., becoming more similar to an interlocutor) indicates social closeness and facilitates communication. Moreover, factors like perceived social distance/friendliness/competence of the interlocutor have been demonstrated to highly influence the degree of interlocutors' acoustic similarity (Schweitzer & Lewandowski, 2014). However, it remains unclear how factors such as social distance and a particular pragmatic act (e.g., apology) interactively shape accommodation behavior. Thus, in the present study, I explored how vocal accommodation in conflicting situations can be explained by socio-pragmatic factors such as interpersonal relationships (i.e., politeness) and pragmatic events occurring in apology-centered interactions. I analyzed 28 dyadic conversations where 14 target speakers (8 males and 6 females) interacted in an apology-centered role-play with a high-status superior or a friend interlocutor.

Results indicate that the conversations with a high status superior are characterized by more divergence behavior than the conversations with a friend. Furthermore, regarding gender, female speakers' conversations show more divergence or maintenance, while male speakers show more variable accommodation behavior (in fundamental frequency (F0)) with a preference for convergence (in F0 variability) in the friend's conversations. These findings suggest that interpersonal power dynamics between interlocutors could serve as a predictor for accommodation behavior.

Study 6: Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing

In human-human interactions, the situational context plays a significant role in the degree of speakers' accommodation. In this study, I investigate whether the degree of accommodation in a human-robot computer game is affected by (a) the duration of the interaction and (b) the players' success in the game. Thirty teams of two players (10 teams were male-male, 10 female-female, and 10 male-female) played two card games with a conversational robot in which they had to find the correct order of five cards. After game 1, the players received the game's result. Results revealed that (a) the duration of the game did not influence the degree of F0 accommodation and (b) the result of Game 1 correlated with the degree of F0 accommodation in Game 2 (higher success equals lower Euclidean distance). The findings suggest that accommodation between speakers is not necessarily a function of the duration of a conversation, but situational factors, like winning the game, can influence speakers' convergence.

Study 7: Revisiting robot-directed speech effects in spontaneous Human-Human-Robot interactions

In light of the recent advances in social robots, the multiparty human-robot interaction is becoming more and more integrated into our everyday life, which might

affect humans' perception of robots. I question whether a well-known speaking style (robot-directed speech) can be observed in a naturalistic setting. In this study, I investigate the differences between human-directed and robot-directed speech during spontaneous human-human-robot interactions. The interactions under study are different from previous studies in that it uses a physical social robot, and the robot has a similar role to the human interlocutors, which leads to more spontaneous turn-taking. Twenty conversations were extracted from a multiparty human-robot discussion corpus, where two humans are playing a collaborative card game with a social robot. There were significant differences between human- and robot-directed speech for speaking rate and the total utterance duration. These results align with previous studies on robot-directed speech and confirm that this difference also holds when the conversations are more spontaneous.

Part I

Content-driven adaptation

Chapter 2

The effect of predictability on German stop voicing is phonologically selective

This article was originally published in:

Ibrahim, Omnia; Yuen, Ivan; Andreeva, Bistra; Möbius, Bernd (2022) The effect of predictability on German stop voicing is phonologically selective, Proc. Speech Prosody 2022, Lisbon, Portugal, pages 669-673, doi: 10.21437/SpeechProsody.2022-136.

Abstract

Cross-linguistic evidence suggests that syllables in predictable contexts have shorter duration than in unpredictable contexts. However, it is not clear if predictability uniformly affects phonetic cues of a phonological feature in a segment. The current study explored the effect of syllable-based predictability on the durational correlates of the phonological stop voicing contrast in German, viz. voice onset time (VOT) and closure duration (CD), using data in Ibrahim et al. (2021). The target stop consonants /b, p, d, k/ occurred in stressed CV syllables in polysyllabic words embedded in a sentence, with either voiced or voiceless preceding contexts. The syllable occurred in either a low or a high predictable condition, which was based on a syllable-level trigram language model. We measured VOT and CD of the target consonants (voiced vs. voiceless). Our results showed an interaction effect of predictability and the voicing status of the target consonants on VOT, but a uniform effect on closure duration. This interaction effect on a primary cue like VOT indicates a selective effect of predictability on VOT, but not on CD. This suggests that the effect of predictability is sensitive to the phonological relevance of a language-specific phonetic cue.

Chapter 3

Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles

This article was originally published in:

Omnia Ibrahim, Homa Asadi, Eman Kassem, Volker Dellwo (2020), Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles, Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, pages 5337–5342, https://aclanthology.org/2020.lrec-1.657.

Abstract

Databases for studying speech rhythm and tempo exist for numerous languages. The present corpus was built to allow comparisons between Arabic speech rhythm and other languages. 10 Egyptian speakers (gender-balanced) produced speech in two different speaking styles (read and spontaneous). The design of the reading task replicates the methodology used in the creation of BonnTempo corpus (BTC). During the spontaneous task, speakers talked freely for more than one minute about their daily life and/or their studies, then they described the directions to come to the university from a famous near location using a map as a visual stimulus. For corpus annotation, the database has been manually and automatically time-labeled, which makes it feasible to perform a quantitative analysis of the rhythm of Arabic in both Modern Standard Arabic (MSA) and Egyptian dialect variety. The database serves as a phonetic resource, which allows researchers to examine various aspects of Arabic supra-segmental features and it can be used for forensic phonetic research, for comparison of different speakers, analyzing variability in different speaking styles, and automatic speech and speaker recognition.

Part II

Acoustic noise-driven adaptation

Chapter 4

The effect of Lombard speech modifications in different information density contexts

This article was originally published in:
Ibrahim, Omnia; Yuen, Ivan; van Os, Marjolein; Andreeva, Bistra; Möbius, Bernd (2021), The effect of Lombard speech modifications in different information density contexts, 32nd conference on electronic speech signal processing (ESSV2021), Berlin, Germany, Pages 185-191.

Abstract

Speakers adapt their speech to increase clarity in the presence of background noise (Lombard speech) (Lu & Cooke, 2009a; Brumm & Zollinger, 2011). However, they also modify their speech to be efficient by shortening word duration in more predictable contexts (Aylett & Turk, 2006). To meet these two communicative functions, speakers will attempt to resolve any conflicting communicative demands. The present study focuses on how this can be resolved in the acoustic domain. A total of 1520 target CV syllables were annotated and analysed from 38 German speakers in 2 white-noise (no noise vs. -10 dB SNR) and 2 surprisal (H vs. L) contexts. Median fundamental frequency (F0), intensity range, and syllable duration were extracted. Our results revealed effects of both noise and surprisal on syllable duration and intensity range, but only an effect of noise on F0. This might suggest redundant (multi-dimensional) acoustic coding in Lombard speech modification, but not so in surprisal modification.

Chapter 5

The combined effects of contextual predictability and noise on the acoustic realization of German syllables

This article was originally published in:

Ibrahim, Omnia; Yuen, Ivan; van Os, Marjolein; Andreeva, Bistra; Möbius, Bernd (2022) The combined effects of contextual predictability and noise on the acoustic realisation of German syllables, Journal of the Acoustical Society of America, Pages 911-920, doi: 10.1121/10.0013413

Abstract

Speakers tend to speak clearly in noisy environments, while they tend to reserve effort by shortening word duration in predictable contexts. It is unclear how these two communicative demands are met. The current study investigates the acoustic realizations of syllables in predictable vs. unpredictable contexts across different background noise levels. Thirty-eight German native speakers produced 60 CV syllables in two predictability contexts in three noise conditions (reference = quiet, 0 dB and -10 dB SNR). Duration, intensity (average and range), F0 (median), and vowel formants of the target syllables were analysed. The presence of noise yielded significantly longer duration, higher average intensity, larger intensity range and higher F0. Noise levels affected intensity (average and range) and F0. Low predictability syllables exhibited longer duration and larger intensity range. However, no interaction was found between noise and predictability. This suggests that noise-related modifications might be independent from predictability-related changes, with implications for including channel-based and message-based formulations in speech production.

Part III

Listener-driven adaptation

Chapter 6

Within-speaker accommodation behavior in apology-centered interactions: The role of socio-pragmatic factors

This article was originally published in:
Ibrahim, Omnia; Hübscher, Iris (2023) Within-speaker accommodation behavior in apology-centered interactions: the role of socio-pragmatic factors. In Pragmatics & Beyond New Series (Multimodal Im/politeness: Signed, spoken, written), ed. by Andreas H. Jucker, Iris Hübscher and Lucien Brown, Pages 185–211, doi: 10.1075/pbns.333.07ibr.

Abstract

In the present study, we explore the extent to which vocal accommodation in conflicting situations can be explained by socio-pragmatic factors such as interpersonal relationships (i.e. politeness) and pragmatic events occurring in apology-centered interactions. We analyzed 28 dyadic conversations where 14 target speakers (8 males and 6 females) interacted in apology-centered role-plays with a status superior and a friend interlocutor separately. The individual utterances of the conversations were manually annotated phonetically, orthographically, and pragmatically. Accommodation was measured with the difference-in-distance paradigm comparing the first and last 30% of the conversations. Results indicate that the conversations with a status superior are characterized by more divergence behavior than the conversations with a friend. Furthermore, in regard to gender, female speakers' conversations show more divergence or maintenance while male speakers show more variable accommodation behavior (in fundamental frequency (F0)) with a preference for convergence (in F0 variability) in the friend's conversations. These findings suggest that interpersonal

power dynamics between interlocutors could serve as a predictor for accommodation behavior.

Chapter 7

Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing

This article was originally published in:
Ibrahim, Omnia; Skantze, Gabriel; Stoll, Sabine; Dellwo, Volker (2019) Fundamental Frequency Accommodation in Multi-Party Human-Robot Game Interactions: The Effect of Winning or Losing. Proc. Interspeech 2019, Pages 3980-3984, doi: 10.21437/Interspeech.2019-2496.

Abstract

In human-human interactions, the situational context plays a large role in the degree of speakers' accommodation. In this paper, we investigate whether the degree of accommodation in a human-robot computer game is affected by (a) the duration of the interaction and (b) the success of the players in the game. 30 teams of two players played two card games with a conversational robot in which they had to find a correct order of five cards. After game 1, the players received the result of the game on a success scale from 1 (lowest success) to 5 (highest). Speakers' fo accommodation was measured as the Euclidean distance between the human speakers and each human and the robot. Results revealed that (a) the duration of the game had no influence on the degree of fo accommodation and (b) the result of Game 1 correlated with the degree of fo accommodation in Game 2 (higher success equals lower Euclidean distance). We argue that game success is most likely considered as a sign of the success of players' cooperation during the discussion, which leads to a higher accommodation behavior in speech.

Chapter 8

Revisiting robot directed speech effects in spontaneous Human-Human-Robot interactions

This article was originally published in:
Ibrahim, Omnia; Skantze, Gabriel (2021) Revisiting robot directed speech effects in spontaneous human-human-robot interactions, Proc. conference on Human Perspectives on Spoken Human-Machine Interaction, Frankfurt (Virtually), pages 6-10, doi : 10.6094/UNIFR/223822.

Abstract

In this paper, we investigate the differences between human-directed speech and robot-directed speech during spontaneous human-human-robot interactions. The interactions under study are different from previous studies, in the sense that the robot has a more similar role as the human interlocutors, which leads to more spontaneous turn-taking. 20 conversations were extracted from a multi-party human-robot discussion corpus, where two humans are playing a collaborative card game with a social robot. Each utterance in the conversations was manually labeled according to addressee (robot or human). The following acoustic features were extracted: fundamental frequency, intensity, speaking rate, and total utterance duration. There were significant differences between human- and robot-directed speech for speaking rate and the total utterance duration. These results are in line with previous studies on robot-directed speech, and confirms that this difference holds also when the conversations are of a more spontaneous nature.

Chapter 9

Future directions

This Ph.D. thesis gives insights on speaker adaptation as a function of I) message encoding, II) channel encoding, and III) listener encoding, and its findings have implications for including channel-based and listener-based formulations in speech production models along with message-based formulations. However, several questions remain to be answered. Therefore, this chapter describes some future directions.

9.1 The interplay between predictability and speech tempo effects on Modern Standard Arabic

Background: Predictability has pervasive effect on the acoustic realization of speech (Aylett & Turk, 2004; Frank & Jaeger, 2008; Crocker et al., 2016). Generally, duration is shortened in more predictable contexts and conversely lengthened in less predictable contexts. In this study, I will explore whether the predictability effect, as previously observed for most languages like English, is also found in Arabic, which has relatively rich inflectional morphology. Arabic has a non-concatenative (inflectional) morphology. Word forms (e.g., [katab] write) result from the interleaving of a consonantal root "ktb" conveying semantics of the word, and a vocalic word pattern "faAal" conveying morphosyntactic information.

Further work is needed to establish whether the predictability effects come from word frequency or the root/pattern frequency in an inflectional language like Arabic. On the other hand, changing the speech tempo will affect the segments' duration. Fast speech tempo will result in reduction or even deletion of some segments.

Research question: Is there any interaction between word (or stem) frequency and speech tempo on the acoustic realization of the vowels and consonants?

To answer this question, I will use the modern standard Arabic (MSA) corpus collected in study 1. The short story was recorded in 5 different tempos, from very

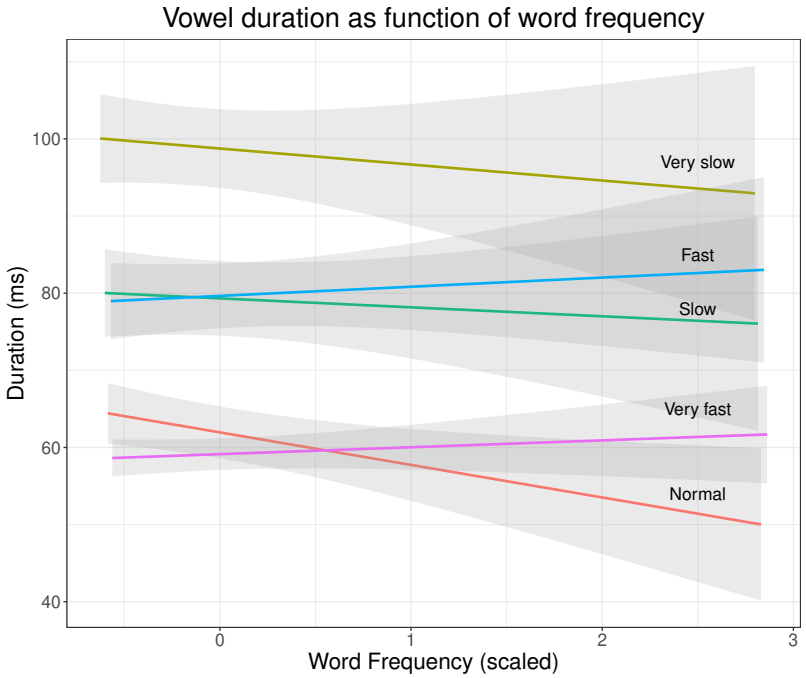

Figure 9.1: Preliminary analysis of vowel duration of the MSA short story in the 5 tempo. Data was extracted from one speaker.

slow to very fast. Word and stem frequencies will be extracted from Aralex[1], an MSA database based on 40 million words and contains information about the word and stem frequencies along with bi- and tri-grams in orthographic forms (Boudelaa & Marslen-Wilson, 2010).

From a preliminary analysis of one speaker (Figure 9.1), I found a relation between vowel duration and carrier word frequency; vowels in high-frequent words are shorter than in less frequent words. However, this effect is mitigated or even reversed in fast (& very fast) speaking rate. Is this idiosyncratic of this speaker or a general trend? Generally, I predict that speakers will not reduce all vowels' duration in the same way (uniformly); vowels in frequent words will be shorter than those in less frequent words. Moreover, in the fastest speaking rate, vowels in frequent words will additionally have extra reductions.

9.2 The relationship between predictability and accommodation behavior

Background: In this thesis, I have explored accommodation phenomena from a social perspective investigating the effects of situational and pragmatic factors on speakers' accommodation behaviour. However, studies have also reported frequency effects on convergence (in voice onset time [VOT] (Nielsen, 2011), in vowel formants

[1]https://aralex.mrc-cbu.cam.ac.uk/aralex.online/login.jsp

(Babel, 2010), in AXB (Dias & Rosenblum, 2016)). Goldinger (1998) found that low-frequency words elicited greater phonetic convergence than high-frequency words, which was replicated in Goldinger & Azuma (2004).

One potential explanation for the observed effect is that frequent syllables have been argued to be stored as a holistic phonetic motor plan/unit for ease of retrieval, relative to less frequent syllables, which are computed on-line (Whiteside & Varley, 1998; Bürki et al., 2015; Laganaro, 2019). This on-line computing of less frequent syllables facilitates more convergence due to the greater influence of the interlocutor's voice during conversation. Frequency is considered a global measure of predictability; what about local contextual predictability?

Research question: Will the speaker converge mainly on less predicted units? Do patterns change during interaction?

Following the frequency effect, I hypothesize that less predicted units will evoke greater convergence than highly predictable units.

9.3 What makes a good counsellor? Using phonetic accommodation to distinguish between high-quality and low-quality counselling conversations

Background: Nowadays, an increasing number of people suffer from behavioral health problems, including uncontrolled smoking and drinking. Therefore, unsuccessful interaction between the therapist and the patient would dramatically affect the overall treatment. Therefore, methods are needed to assess the dialogue quality and better understand what makes good counseling beyond user-self subjective Likert scale rating.

Accommodation has been proposed as a strategy to signal and manage the social distance between speakers (Giles et al., 1991). Studies have shown that convergent vocal accommodation (i.e., becoming more similar to the interlocutor) indicates social closeness and facilitates communication. Moreover, factors like perceived social distance/friendliness/competence of the interlocutor have been demonstrated to highly influence the degree of the interlocutors' acoustic similarity Schweitzer & Lewandowski (2014). In this research, I will investigate the potential of using linguistic accommodation in a clinical setting to evaluate the quality of counseling conversations.

Research question: Could we use phonetic accommodation to distinguish between high-quality and low-quality behavioral counseling conversations? [2]

[2]The proposed idea was presented and discussed in language and medicine colloquium, university of Zürich, 2021.

20 good and 20 bad counseling conversations will be selected from Pérez-Rosas et al. (2019). They are of 6 to 20 minutes duration. The sessions cover health topics: smoking, alcohol consumption, substance abuse, etc. I hypothesize that high-quality counseling conversations will show higher level of phonetic convergence between the therapist and the patient, while low-quality conversations will show divergence or maintenance.

Bibliography

Arantes, P., & Eriksson, A. (2019). Quantifying Fundamental Frequency Modulation as a Function of Language, Speaking Style and Speaker. In *Proceedings of Interspeech* (pp. 1716–1720). doi:doi: 10.21437/Interspeech.2019-2857.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*, 31–56. doi:doi: 10.1177/00238309040470010201.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*, 3048–3058. doi:doi: 10.1121/1.2188331.

Babel, M. (2010). Dialect divergence and convergence in new zealand english. *Language in Society*, *39*, 437–456. doi:doi: 10.1017/S0047404510000400.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*, 177–189. doi:doi: 10.1016/j.wocn.2011.09.001.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS* (p. 833–836).

Boril, H., & Pollák, P. (2005). Design and collection of czech lombard speech database. In *Proceedings of Interspeech* (pp. 1577–1580). doi:doi: 10.21437/Interspeech.2005-461.

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, *42*, 481–487. doi:doi: 10.3758/BRM.42.2.481.

Bradlow, A. (2002). Confluent talker- and listener-related forces in clear speech production. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 241–273). Mouton de Gruyter.

Brandt, E., Andreeva, B., & Möbius, B. (2019). Information density and vowel dispersion in the productions of Bulgarian l2 speakers of German. In *Proceedings of ICPhS* (pp. 3165–3169). Melbourne. URL: http://www.sfb1102.uni-saarland.de/wp/wp-content/uploads/2019/09/brandt_etal_icphs2019.pdf.

Brumm, H., & Zollinger, S. A. (2011). The evolution of the lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*, 1173–1198. URL: http://www.jstor.org/stable/41445240.

Buz, E., & Jaeger, T. F. (2016). The (in)dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*, *31*, 404–424. doi:doi: 10.1080/23273798.2015.1105984.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290. doi:doi: 10.1017/S0954394502143018.

Bürki, A., Cheneval, P. P., & Laganaro, M. (2015). Do speakers have access to a mental syllabary? ERP comparison of high frequency and novel syllable production. *Brain Lang*, *150*, 90–102. doi:doi: 10.1016/j.bandl.2015.08.006.

Castellanos, A., Benedí, J.-M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect. *Speech Communication*, *20*, 23 – 35. doi:doi: 10.1016/S0167-6393(96)00042-8.

Cohen Priva, U., & Sanker, C. (2019). Limitations of difference-in-difference for measuring convergence. *Journal of the Association for Laboratory Phonology*, *10*, 1–29. doi:doi: 10.5334/labphon.200.

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech Language*, *28*, 543–571. doi:doi: 10.1016/j.csl.2013.08.003.

Crocker, M. W., Demberg, V., & Teich, E. (2016). Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, *30*, 77–81.

Demberg, V., Sayeed, A., Gorinski, P., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 356–367). Jeju Island, Korea: Association for Computational Linguistics. URL: https://aclanthology.org/D12-1033.

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics*, *78*, 317–333. doi:doi: 10.3758/s13414-015-0982-6.

Englund, K. T. (2017). Hypoarticulation in infant-directed speech. *Applied Psycholinguistics*, *39*, 67–87. doi:doi: 10.1017/S0142716417000480.

Frank, A. F., & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. URL: https://escholarship.org/uc/item/7d08h6j4.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*, 789–806. doi:doi: 10.1016/j.jml.2011.11.006.

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech Language*, *28*, 580 – 597. doi:doi: 10.1016/j.csl.2013.07.005.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics* Studies in Emotion and Social Interaction (p. 1–68). Cambridge University Press. doi:doi: 10.1017/CBO9780511663673.001.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological review*, *105*, 251–279. doi:doi: 10.1037/0033-295x.105.2.251.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic bulletin & review*, *11*, 716–722. doi:doi: 10.3758/BF03196625.

Gregory Jr., S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology*, *70*, 1231–1240. doi:doi: 10.1037/0022-3514.70.6.1231.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. URL: https://aclanthology.org/N01-1021.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, (pp. 1–16). doi:doi: 10.1111/lnc4.12196.

Hansen, J. H. L., Lee, J., Ali, H., & Saba, J. N. (2020). A speech perturbation strategy based on "lombard effect" for enhanced intelligibility for cochlear implant listeners. *The Journal of the Acoustical Society of America*, *147*, 1418–1428. doi:doi: 10.1121/10.0000690.

Heath, J. (2017). How automatic is convergence? evidence from working memory. *Proceedings of the Linguistic Society of America*, *2*, 35. doi:doi: 10.3765/plsa.v2i0.4088.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190. doi:doi: 10.1080/17470218.2013.850521.

Ibrahim, O., Yuen, I., van Os, M., Andreeva, B., & Möbius, B. (2021). The effect of lombard speech modifications in different information density contexts. In *Elektronische Sprachsignalverarbeitung, Tagungsband der 32. Konferenz (Berlin)* (pp. 185–191). Dresden: TUDpress.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62. doi:doi: 10.1016/j.cogpsych.2010.02.002.

Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, *20*, 13–22. doi:doi: 10.1016/S0167-6393(96)00041-6.

Kriz, S., Anderson, G., Bugajska, M., & Trafton, J. G. (2009). Robot-directed speech as a means of exploring conceptualizations of robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* HRI '09 (p. 271–272). New York, NY, USA: Association for Computing Machinery. doi:doi: 10.1145/1514095.1514171.

Labov, W. (2006). *The Social Stratification of English in New York City*. (2nd ed.). Cambridge University Press. doi:doi: 10.1017/CBO9780511618208.

Laganaro, M. (2019). Phonetic encoding in utterance production: A review of open issues from 1989 to 2018. *Language and cognition*, *34*, 1193–1201. doi:doi: 10.1080/23273798.2019.1599128.

Lee, C.-C., Katsamanis, A., Black, M. P., Baucom, B. R., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (2014). Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech Language*, *28*, 518–539. doi:doi: 10.1016/j.csl.2012.06.006.

Leongómez, J. D., Pisanski, K., Reby, D., Sauter, D., Lavan, N., Perlman, M., & Varella Valentova, J. (2021). Voice modulation: from origin and mechanism to social impact. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*, 20200386. doi:doi: 10.1098/rstb.2020.0386.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech* (pp. 3081–3084). doi:doi: 10.7916/D8V12D8F.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Maladiers Oreille, Larynx, Nez, Pharynx (Annals of diseases of the ear, larynx, nose and pharynx*, *37*, 101–119.

Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *Journal of the Acoustical Society of America*, *124*, 3261–3275. doi:doi: 10.1121/1.2990705.

Lu, Y., & Cooke, M. (2009a). The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*, 1253 – 1262. doi:doi: 10.1016/j.specom.2009.07.002.

Lu, Y., & Cooke, M. (2009b). Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America*, *126*, 1495–1499. doi:doi: 10.1121/1.3179668.

Michael, T., & Ibrahim, O. (2022). Lexical frequency and listener's response to packet loss in telephone conversations. In *33rd Conference on Electronic Speech Signal Processing (ESSV2022)*.

Michalsky, J., & Schoormann, H. (2017). Pitch Convergence as an Effect of Perceived Attractiveness and Likability. In *Proceedings of Interspeech* (pp. 2253–2256). doi:doi: 10.21437/Interspeech.2017-1520.

Mikeev, Y. V. (1971). Soviet physics-acoustics. *Evolution and Human Behavior*, *16*, 474–477.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*, 1048–1058. doi:doi: 10.1044/1092-4388(2004/078).

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation:: The role of perception. *Journal of Language and Social Psychology*, *21*, 422–432. doi:doi: 10.1177/026192702237958.

Nielsen, K. (2011). Specificity and abstractness of vot imitation. *Journal of Phonetics*, *39*, 132–142. doi:doi: 10.1016/j.wocn.2010.12.007.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*. doi:doi: 10.3389/fpsyg.2013.00559.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*, 2382–2393. doi:doi: 10.1121/1.2178720.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, Psychophysics*, *79*, 637–659. doi:doi: 10.3758/s13414-016-1226-0.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11. doi:doi: 10.1016/j.wocn.2018.04.001.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1 – 17. doi:doi: 10.1016/j.jml.2014.10.003.

Pérez-Rosas, V., Wu, X., Resnicow, K., & Mihalcea, R. (2019). What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 926–935). doi:doi: 10.18653/v1/P19-1088.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190. doi:doi: 10.1017/S0140525X04000056.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329–347. doi:doi: 10.1017/S0140525X12001495.

Pickering, M. J., & Garrod, S. (2021). Understanding dialogue: Language use and social interaction. Cambridge: Cambridge University Press.

Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (2021). A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 949–962). Association for Computational Linguistics.

Rappaport, W. (1958). *Uber Messungen der Tonhohen Verteilung in der deutschen Sprache*.

Rouas, J.-L., Beppu, M., & Adda-Decker, M. (2010). Comparison of spectral properties of read, prepared and casual speech in French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/704_Paper.pdf.

Ruch, H., Zürcher, Y., & Burkart, J. M. (2018). The function and mechanism of vocal accommodation in humans and other primates. *Biological Reviews*, *93*, 996–1013. doi:doi: 10.1111/brv.12382.

Scholz, F., & Zhu, A. (2013). Package 'ksamples' title k-sample rank tests and their combinations. URL: https://CRAN.R-project.org/package=kSamples.

Schweitzer, A., & Lewandowski, N. (2014). Social factors in convergence of f1 and f2 in spontaneous speech. In *The International Seminar on Speech Production*. Cologne. doi:doi: 10.13140/2.1.3709.5689.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 379–423. URL: http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48.

Shannon, C. E., & Weaver, W. (1949). A mathematical theory of communication. *University of Illinois Press*, (pp. 1–29). URL: https://pure.mpg.de/rest/items/item_2383164/component/file_2383163/content.

Shepard, C., Giles, H., & Poired, B. (2001). *Communication Accommodation Theory*. New York: Wiley.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Linguistics Lang. Compass*, *3*, 621–640. doi:doi: 10.1111/j.1749-818X.2009.00125.x.

Torabi Asr, F., & Demberg, V. (2015). Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 118–128). URL: https://aclanthology.org/W15-0117.

Traunmüller, H., & Eriksson, A. (1994). *The frequency range of the voice fundamental in the speech of male and female adults*. Technical Report Department of Linguistics, University of Stockholm.

Turnbull, R. (2019). Listener-oriented phonetic reduction and theory of mind. *Language and Cognition*, *34*, 747–768. doi:doi: 10.1080/23273798.2019.1579349.

Van Os, M., Kray, J., & Demberg, V. (2021). Mishearing as a side effect of rational language comprehension in noise. *Frontiers in Psychology*, *12*. doi:doi: 10.3389/fpsyg.2021.679278.

Whiteside, S., & Varley, R. (1998). A reconceptualisation of apraxia of speech: A synthesis of evidence. *Cortex*, *34*, 221–231. doi:doi: 10.1016/S0010-9452(08)70749-4.

Winter, B., Oh, G. E., Hübscher, I., Idemaru, K., Brown, L., Prieto, P., & Grawunder, S. (2021). Rethinking the frequency code: a meta-analytic review of the role of acoustic body size in communicative phenomena. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*, 20200400. doi:doi: 10.1098/rstb.2020.0400.

Zellers, M., & Schweitzer, A. (2017). An investigation of pitch matching across adjacent turns in a corpus of spontaneous German. In *Proceedings of Interspeech* (pp. 2336–2340). doi:doi: 10.21437/Interspeech.2017-811.